CLAIMS

What is claimed is:

1. A system that facilitates extracting data in connection with spam processing, comprising:

a component that receives a message and extracts a set of features associated with some part, content or content type of a message; and

an analysis component that at least examines consecutiveness of characters within a subject line of the message in connection with building a filter.

2. The system of claim 1, the analysis component determines frequency of consecutive repeating characters within the subject line of the message.

3. The system of claim 2, the characters comprise letters, numbers, or punctuation.

4. The system of claim 1, the analysis component determines frequency of white space characters within the subject line of the message.

5. The system of claim 1, the analysis component determines distance between at least one alpha-numeric character and a blob.

6. The system of claim 1, the analysis component determines a maximum number of consecutive, repeating characters and stores this information.

7. The system of claim 1, the analysis component establishes ranges of consecutive, repeating characters, the ranges corresponding to varying degrees of spaminess, whereby messages can be sorted by their respective individual count of consecutive repeating characters.

8. The system of claim 1, the analysis component further determines content type associated with the message.

9. The system of claim 8, the analysis component compares the content type of a current message to stored content types of a plurality of other messages to facilitate determining whether the message is spam.

10. The system of claim 8, the content type is case-sensitive.

11. The system of claim 8, the content type comprises a primary content-type and a secondary content-type.

12. The system of claim 1, the analysis component further determines time stamps associated with the message.

13. The system of claim 12, the analysis component determining a delta between time stamps.

14. The system of claim 13, the delta is between a first and a last time stamp.

15. The system of claim 1, the analysis component determines at least one of: a percentage of white space to non-white space in the subject line of the message and a percentage of non-white space and non-numeric characters that are not letters in the subject line of the message.

16. The system of claim 1, the filter being a spam filter.

17. The system of claim 1, the filter being a parental control filter.

18. The system of claim 1, further comprising a machine learning system component that employs at least a subset of extracted features to learn at least one of spam and non-spam.

19. A system that facilitates extracting data in connection with spam processing, comprising:

a component that receives an item and extracts a set of features associated with a message; and

an analysis component that determines whether an embedded message or attachment is associated with the message.

20. The system of claim 19, the analysis component identifies a type of embedded message or attachment to facilitate predicting whether the message is spam.

21. The system of claim 19, further comprising a component that employs at least a subset of the extracted features to populate at least one feature list.

22. The system of claim 21, the at least one feature list is any one of a list of good users, a list of spammers, a list of positive features indicating legitimate sender, and a list of features indicating spam.

23. The system of claim 19, further comprising a component that examines at least a portion of a message body.

24. The system of claim 23, the component examines at least a beginning portion of the message body.

25. The system of claim 23, the component determines at least one of: a percentage of white space to non-white space in the message body and a percentage of non-white space and non-numeric characters that are not letters in the message body.

26. The system of claim 23, the component determines a percentage or a number of consecutive lines of a message body to examine.

27. The system of claim 23, the component examines the message body for the presence of at least one blob or consecutive, repeating characters.

28. The system of claim 27, the characters comprising letters, punctuation, and numbers.

29. A method that facilitates spam detection and prevention comprising:

receiving a plurality of messages, the plurality comprising at least a first and a second message;

extracting at least a subset of information from the plurality of messages, the information being from at least one of a subject line, a content-type header, a received header, and a message body; and

analyzing the subset of information to generate one or more features to facilitate training a filter.

30. The method of claim 29, analyzing the subset of information comprises determining a number of consecutive repeating characters within the subject line or the message body of the message.

31. The method of claim 30, the characters comprise letters, numbers, or punctuation.

32. The method of claim 29, analyzing the subset of information comprises determining a frequency of white space characters within the subject line of the message.

33. The method of claim 29, analyzing the subset of information comprises determining a distance between at least one alpha-numeric character and a blob.

34. The method of claim 29, analyzing the subset of information comprises:

determining a maximum number of consecutive, repeating characters and storing this information; and

establishing ranges of consecutive, repeating characters, the ranges corresponding to varying degrees of spaminess, whereby messages can be sorted by their respective individual count of consecutive repeating characters.

35. The method of claim 29, analyzing the subset of information comprises:

determining content type associated with the message; and

comparing the content type of a current message to stored content types of a plurality of other messages to facilitate determining whether the message is spam.

36. The method of claim 29, analyzing the subset of information comprises:

determining time stamps associated with the message; and

determining a delta between a first time stamp and a last time stamp, the first time stamp being located in a Received header and the last time stamp being located in a Date header at the message's destination.

37. The method of claim 29, analyzing the subset of information comprises determining a percentage or a number of consecutive lines of a message body to examine at least one of: a percentage of white space to non-white space in the subject line of the message and a percentage of non-white space and non-numeric characters that are not letters in the subject line of the message.

38. The method of claim 29, analyzing the subset of information comprises determining whether an embedded message or an attachment exists in the message and identifying a type of embedded message or attachment to facilitate predicting whether the message is spam.

39. The method of claim 29, analyzing the subset of information comprises examining at least a beginning portion of the message body.

40. A computer-readable medium having stored thereon the following computer executable components:

a component that receives a message and extracts a set of features associated with some part, content or content type of a message;

an analysis component that examines at least consecutiveness of characters within a subject line of the message in connection with building a filter;

a component that determines whether an embedded message or attachment is associated with the message; and

a component that determines a percentage or a number of consecutive lines of a message body to examine and that examines the message body for the presence of at least one blob or consecutive, repeating characters.

41. A system that facilitates printing from a web page comprising:

means for receiving a plurality of messages, the plurality comprising at least a first and a second message;

means for extracting at least a subset of information from the plurality of messages, the information being from at least one of a subject line, a content-type header, a received header, and a message body; and

means for analyzing the subset of information to generate one or more features to facilitate training a filter, the means for analyzing the subset of information comprising:

means for determining a number of consecutive repeating characters within the subject line or the message body of the message;

means for determining a delta between a first time stamp and a last time stamp associated with the messasge;

means for determining whether an embedded message or an attachment exists in the message and identifying a type of embedded message or attachment to facilitate predicting whether the message is spam; and

means for determining a percentage or a number of consecutive lines of a message body to examine at least one of: a percentage of white space to non-white space in the subject line of the message and a percentage of non-white space and non-numeric characters that are not letters in the subject line of the message.